[IR2018-19-HW1] 1183732 Davide Martini

1. Descrizione del lavoro svolto

Lo sviluppo dell'homework è stato effettuato in ambiente *Linux, Ubuntu 18.04*, tramite l'utilizzo di:

- *terrier-core-4.4*.
- *trec_eval 9.0.4*.
- *PyCharm*.

È stato inoltre prodotto uno script in *Python 3.7* per automatizzare l'esecuzione dell'indicizzazione, la produzione delle varie run, il calcolo delle misure, produrre il test statistico ANOVA 1-way, il Tukey HSD test e per ottenere i plot finali. Lo script è contenuto nella repository con il nome di *ir_script.py*. Di seguito verrà riportata la struttura del codice con una breve spiegazione del suo contenuto.

Per automatizzare l'esecuzione di tutto il workflow necessario per ottenere i risultati, il codice esegue i seguenti passi:

I. Viene impostata una cartella principale dove è contenuta la collezione di documenti, gli eseguibili dei software *terrier* e *trec_eval* e dove verranno salvati i risultati ottenuti.
II. Viene inizializzato il software terrier e modificate le proprietà per l'indicizzazione e per l'esecuzione delle run. Questo passaggio viene ripetuto per tutti i modelli da analizzare. Sono stati salvati i vari file degli indici nella cartella '*indexes/'* e i file delle run nella cartella '*indexes/run/*'.
III. Viene eseguita la valutazione tramite il software *trec_eval* sui i file risultanti delle run. I file contenenti le misure per la valutazione sono stati successivamente salvati nella cartella '*indexes/run/eval/*'.
IV. È stata creata una struttura per contenere le varie misure di valutazione ottenute dal software *trec_eval*, in modo da avere un facile reperimento di queste per i test statistici e per i successivi plot.
V. Sono stati creati dei file, per ogni misura, in cui ogni colonna rappresenta una run.
VI. È stata successivamente effettuata l'ANOVA 1-way ed il Tukey HSD test.
VII. Sono stati infine prodotti i plot per ogni topic e run delle misure *P(10)* e *Rprec*. Un grafico per ogni run per la *MAP*.

2. Risultati ottenuti

Il numero di documenti indicizzati presenti all'interno della collezione è pari a 528155.

È stato sviluppato il test statistico ANOVA 1-way, Tukey HSD pairwise test e Tukey HSD multiple comparisons. L'obiettivo del test statistico ANOVA 1-way è capire se i vari modelli analizzati avessero oppure no la stessa media. Nello sviluppo del test, il valore di soglia $\alpha = 0.05$ permette di rifiutare o meno l'ipotesi che i 4 modelli abbiamo la stessa media. I valori ottenuti sono presenti nel file '*run/plot/anova.txt*':

- $F_{stat} = 0.09965024788515882$.
- $p = P\,[F \geq F_{stat} \mid H_0] = 0.9601256241314807$.

Si nota quindi che l'ipotesi non viene rifiutata e quindi i 4 modelli hanno la stessa media.

Per quanto riguarda il Tukey HSD test, i risultati ottenuti per i 4 modelli sono riportati nella *Tabella 1*.

I vari modelli sono riportati con una sigla:

- BM25: modello BM25 con Stopword e Porter Stemmer.
- BM25_stem: modello BM25 senza Stopword, con Porter Stemmer.
- TF_IDF: modello TF_IDF con Stopword e Porter Stemmer.
- TF_IDF_not: modello BM25 senza Stopword e Porter Stemmer.

Multiple Comparison of Means - Tukey HSD, FWER=0.05					
group 1	*group 2*	*meandiff*	*lower*	*upper*	*reject*
BM25	BM25_stem	0.0029	-0.0816	0.0874	False
BM25	TD_IDF_not	-0.0135	-0.098	0.071	False
BM25	TF_IDF	-0.0006	-0.0851	0.0838	False
BM25_stem	TD_IDF_not	-0.0164	-0.1008	0.0681	False
BM25_stem	TF_IDF	-0.0035	-0.088	0.081	False
TD_IDF_not	TF_IDF	0.0128	-0.0716	0.0973	False

Dal Tukey HSD test si vede ogni coppia di modelli è simile e c'è sempre intersezione nell'intervallo di confidenza di un altro modello. Per avere una visione d'insieme è stato effettuato un confronto multiplo tra i vari modelli riportato in *Figura 1* e nel file '*run/plot/TukeyHSDtest.svg*'.

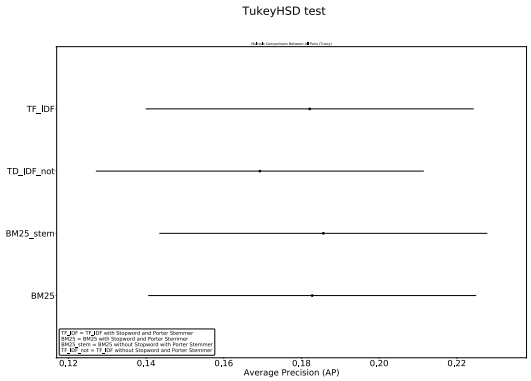

Figura 1. *Tukey HSD test - Multiple Comparisons.*

I modelli quindi sono molto simili, anche se l'indicizzazione del modello TF_IDF_not è risultata più lunga in quanto sono state tenute in considerazione tutte le parole e queste non sono state sottoposte al processo di stemming.
Per quanto riguarda l'analisi della misura *Rprec* il confronto dei grafici, presenti nella cartella '*run/plot/*' con nome '*RprecMODEL.svg*', dove *MODEL* indica il modello considerato, indica una lieve differenza tra i modelli *BM25*, *BM25_stem*, *TF_IDF*, ma evidenzia un forte calo nel modello *TF_IDF_not*. Questa differenza non è presente in tutti i topic ma è abbastanza accentuata nei topic doove si verifica.
I valori ottenuti per la misura *P(10)*, presenti i grafici nella cartella '*run/plot*' con nome '*P_10MODEL.svg*' dove *MODEL* indica il modello considerato, mostrano un andamento abbastanza simile per i modelli *BM25*, *BM25_stem*, *TF_IDF*, ma il modello *TF_IDF* ha un andamento simile per alcuni topic e un valore minore della *Precision* per gli altri.
Da queste due misure, anche se il test di Tukey non ha mostrato una grande differenze, si può definire il modello *TF_IDF_not*, come il peggiore dei 4 per quanto riguarda l'aspetto di efficacia nel reperimento dei documenti.

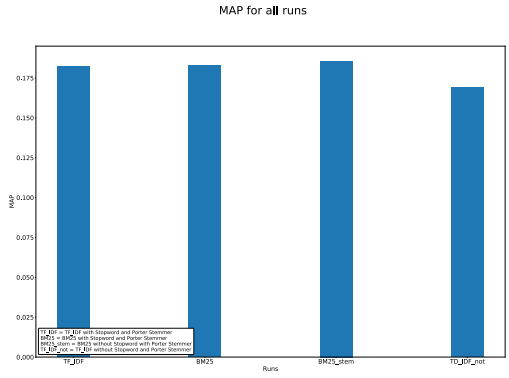

Figura 2. *MAP.*

In *Figura 2*, viene riportato il valore della *MAP* per ogni modello. Notiamo che il modello *TD_IDF_not* ha una *MAP* più bassa tra tutti mentre, come nelle considerazioni precedenti, gli altri 3 modelli assumono valori molto simili.
Nei vari test effettuati, si può vedere come il modello senza Stopword e Porter Stemmer sia il peggiore anche se non di molto rispetto agli altri. Il maggiore tempo di indicizzazione però può portare alla scelta di un altro modello. Nel nostro caso sono stati indicizzati 528155 documenti ma se si considerano collezioni di dimensioni maggiori, questo rallentamento potrebbe influire negativamente nelle prestazioni del sistema.

3. Riferimenti

[1] Github: https://github.com/davidemartini/Information-Retrieval
[2] Trec_eval: https://github.com/usnistgov/trec_eval
[3] Terrier: http://terrier.org/